International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume 2, Issue 3 (March '2016)

A SURVEY ON EXAMINES THE PREDICTION DATA ANALYSIS SYSTEMS IN DATA MINING

P. Kowsalyadevi,
Research Scholar,
Periyar University,
Salem,Tamilnadu,India.

Dr. K.Thangadurai,
Assistant Professor and Head,
Department of Computer Science,
Government Arts College,
Karur,Tamilnadu,India.

Abstract: Predictive data mining automatically create classification model from training dataset and apply such model to automatically predict other classes of unclassified datasets. Predictive data mining deals with learning models to support clinicians in diagnostics, therapeutic, or monitoring tasks. It learns from past experience and applies knowledge gained to future situations, by applying machines learning methods to build multivariate models from different kind of data and subsequently make inferences on unknown data. Machine learning model is related to the exploitation of supervised classification approaches. Prior to applying the learning model, the data is pre processed to remove noise and ensure data mining principle is applied on real data. In Machine Learning and Data Mining fields, classification is usually approached as a supervised learning task. A search algorithm is used to induce a classifier from a set of correctly classified data instances, called the training set. Another set of correctly classified data instances, known as the testing set, is used to measure the quality of the classifier obtained after the learning process. Different kinds of models can be used to represent classifiers, and there are a great variety of algorithms available for inducing classifiers from data. In this paper discussed about previous research related to prediction and classification algorithms used in our work.

Keywords: Data mining, Predictive, Machine Learning, Classification Algorithms

I.INTRODUCTION

Data mining is a multidisciplinary area in which several computing paradigms converge: decision tree construction, rule induction, artificial neural networks, instance-based learning, Bayesian learning, logic programming, statistical algorithms, etc. And some of the most useful data mining tasks and methods are: statistics, visualization, clustering, classification, and association rule mining. These methods uncover new, interesting and useful knowledge based on student usage data [1]. The purpose of Predictive mining model is mainly to predict the future outcome than current behavior. The prediction output can be numeric value or in categorized form. The predictive models are the supervised learning functions which predict the target value. Predictive methods to be utilized in the next phase of empirical study. Use some variables to predict unknown or future values of other variables in addition comparison between these supervised approaches was also conducted to get some insight about the strength and weaknesses of each approach since one of the aims of the study was to determine whether these methods were well suited for extracting the required knowledge. As a result, the predictive method will be able to predict in which cluster the future student will falls into based on the enrollment information [2].

II. LITERATURE SURVEY

Basma Boukenze1, *et al.* [3] in this work they used a learning algorithm C4.5 to predict patients with chronic kidney failure disease (ckd), and patients who are not suffering from this disease (notckd). The classifier used proved its performance in predicting with best results in terms of accuracy and minimum execution time. In this work, they applied C4.5, a learning algorithm that will make classification and prediction on a database to extract knowledge and classify patients into two categories: chronic kidney disease (ckd) and not chronic kidney disease (notckd). In this research, they did use the Waikato Environment for Knowledge Analysis (Weka). It is a comprehensive suite of Java class libraries that implement many algorithms for data mining clustering, classification, regression, analysis of results. This platform gives to researchers a perfect environment to implement and evaluate their classification model comparing to TANAGRA or ORANGE.

Parneet Kaura, *et al.* [4] this researcher focused on identifying the slow learners among students and displaying it by a predictive data mining model using classification based algorithms. They, used classification techniques for prediction on the dataset of 152 students, to predict and analyze student's performance as well slow learners among them. In this work, a model was developed based on some selected student related input variables collected from real world (high schools). Among all data mining classifiers Multi Layer Perception performs best with 75% accuracy and therefore MLP proves to be potentially effective and efficient classifier algorithm. Also comparison of all 5 classifiers with the help of WEKA experimenter is also done, in this case also MLP proves to be best with F-measure of 82%. Therefore, performance of MLP is relatively higher than other classifiers. A model performance chart is also plotted. This research help the institutions to identify students who are slow learners which further provide base for deciding special aid to them.

Irina Ionita, *et al.* [5] This researcher aims at analyzing the possibilities of applying predictive data mining models (decision trees, logistic regression) in order to predict the military career choice among young people. The results indicated that, based on data mining models, there can be created a candidate profile for such a career. The experiments were conducted on a sample data (274 records) and the following algorithms have been applied: logistic regression, J48, JRIP, LMT, REPTree and Simple CART. After analyzing the results, it was observed that the algorithm with the best rate of classification is Simple CART, followed by J48, LMT and logistic regression.

B.Venkatalakshmi, *et al.* [6] this researcher intends to design and develop diagnosis and prediction system for heart diseases based on predictive mining. Number of experiments has been conducted to compare the performance of various predictive data mining techniques including Decision tree and Naïve Bayes algorithms. In this research work, a 13 attribute structured clinical database from UCI Machine Learning Repository has been used as a source data. Decision tree and Naive Bayes have been applied and their performance on diagnosis has been compared. Naïve Bayes outperforms when compared to Decision tree. In this research results many sessions of experiments were conducted with the same datasets in Weka 3.6.0 tool. Data set of 294 records with 13 attributes is used and the outcome reveals that the Naïve Bayes outperforms and sometime Decision Tree. In Future Genetic algorithm will be used in order to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

Soo-Yeon Ji, *et al.* [7] predicted the hourly rainfall in any geographical regions time efficiently. The chance of rain is first determined. Then only if there is any chance of rainfall, the hourly rainfall prediction is performed. Although quite a lot methodology have been introduced to predict hourly prediction, most of them have performance limitations because of the existence of wide range of variation in data and limited amount of data. CART and C4.5 are used to provide outcomes, which may provide hidden and important patterns with transparent reasons. About 18 variables were used from weather station. For validation purpose, 10 fold cross validation method is performed. CART gives slightly better performance than C4.5. Considering the chances, only a small number of instances are left for prediction which makes it hard to predict.

M. Durairaj, *et al.* [8] illustrates a hybrid prediction system consists of Rough Set Theory (RST) and Artificial Neural Network (ANN) for dispensation medical data. The process of developing a new data mining technique and software to assist competent solutions for medical data analysis has been explained. Propose a hybrid tool that incorporates RST and ANN to make proficient data analysis and indicative predictions. The experiments onspermatological data set for predicting excellence of animal semen is carried out. The projected hybrid prediction system is applied for pre-processing of medical database and to train the ANN for production prediction. The prediction accuracy is observed by comparing observed and predicted cleavage rate[20].

Hamidah Jantan, *et al.* [9] they used decision tree C4.5 classification algorithm to generate the classification rules for human talent performance records. In this research, as we can see from result analysis, C4.5 classifier has a great potential for performance prediction. The generated classification rules can be used to predict the performance of an employee whether he/she has potential to be promoted or not, based on his/her performance. Currently, this research is at the stage of system development where the classification rules need to be embedded into the decision support system which is known as Intelligent Decision Support System (IDSS). Finally, the generated rules are evaluated using the unseen data in order to estimate the accuracy of the prediction result.

R. Maciejewski, *et al.* [10] proposed a model for spatiotemporal data, as analysts are searching for regions of space and time with unusually high incidences of events (hotspots), created a predictive visual analytics toolkit that provides analysts with linked spatiotemporal and statistical analytic views. The system models spatiotemporal events through the combination of kernel density estimation for event distribution and seasonal trend decomposition by loss smoothing for temporal predictions.

E. G. Petre, *et al.* [11] presented a small application of CART decision tree algorithm for weather prediction. The data collected is registered over Hong Kong. The data is recorded between 2002 and 2005. The data used for creating the dataset includes parameters year, month, average pressure, relative humidity, clouds quantity, precipitation and average temperature. WEKA, open source data mining software, is used for the implementation of CART decision tree algorithm. The decision tree, results and statistical information about the data are used to generate the decision model for prediction of weather. The way the data is stored about past events is highlighted. The data transformation is required according to the decision tree algorithm in order to be used by WEKA efficiently for weather prediction.

Gerben W. Dekker, *et al.* [12] They described the results of the educational data mining aimed at predicting the Electrical Engineering (EE) students drop out after the first semester of their studies or even before they enter the study program as well as identifying success-factors specific to the EE program. In this research, they demonstrated the effectiveness of several classification techniques and the cost-sensitive learning approach on the dataset from the Electrical Engineering department of Eindhoven University of Technology. The in depth model evaluation pointed to three major improvements that can be assessed. Firstly, a key improvement in this dataset would be to find a solution for the changing course organization over the set. Aggregating the available information about student performance for a course in a way that can be used for all students in the dataset might prevent the type of misclassifications that is now strongly prevalent. A second, related improvement would be a better way to encode grades in general. Mapping all unknown or not available information to zero showed to be not effective. Specifically, Linear Algebra Educational Data Mining grades should be available. A more advanced

solution dealing with missing values also can be considered in this respect. The quality of the classification criterion is the third improvement that might be considered.

Z. Huang, *et al.* [13] applied predictive analytics techniques to establish a decision support system for complex network operation management and help operators predict potential network failures and adapt the network in response to adverse situations. The resultant decision support system enables continuous monitoring of network performance and turns large amounts of data into actionable information.

R. M. Riensche, *et al.* [14] described a methodology and architecture to support the development of games in a predictive analytics context, designed to gather input knowledge, calculate results of complex predictive technical and social models, and explore those results in an engaging fashion.

Esther Ge, *et al.* [15] a number of possible data mining methods that can be applied to do the lifetime prediction of metallic components and how different sources of service life information could be integrated to form the basis of the lifetime prediction model. Researcher compare a number of data mining methods on the data sets provided by our industry partners and analyze what kind of methods are suitable for what kind of data. Firstly, traditional data mining methods like Naïve Bayes, Decision Tree, Neural Network, SVM and M5 etc were applied to build a number of independent predictors for each data set. The results indicate that the best method for predicting the service life depends on the data set used to train the model. Also analyzed the predicted service life from each data set using certain test cases. The testing shows that in some situations, inconsistent predicted results may be presented by the data mining systems due to using three different data sets (information) for a same test case.

Andrew Kusiak, *et al.* [16] have used data pre-processing, data transformations, and data mining approach to elicit knowledge about the interaction between many of measured parameters and patient survival. Two different data mining algorithms were employed for extracting knowledge in the form of decision rules. Those rules were used by a decision-making algorithm, which predicts survival of new unseen patients. Important parameters identified by data mining were interpreted for their medical significance. They have introduced a new concept in their research work, it has been applied and tested using collected data at four dialysis sites. The approach presented in their paper reduced the cost and effort of selecting patients for clinical studies. Patients can be chosen based on the prediction results and the most significant parameters discovered.

Behrouz Minaei-Bidgoli, *et al.* [17] They intends was to classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system. The research presented here was performed on a part of the latest online educational system developed at MSU, the Learning Online Network with Computer-Assisted Personalized Approach (LON-CAPA). Four classifiers were used to segregate the students. A combination of multiple classifiers leads to a significant accuracy improvement in all 3 cases. Weighing the features and using a genetic algorithm to minimize the error rate improves the prediction accuracy at least 10% in the all cases of 2, 3 and 9-Classes. In cases where the number of features is low, the feature weighting worked much better than feature selection. The successful optimization of student classification in all three cases demonstrates the merits of using the LON-CAPA data to predict the students' final grades based on their features, which are extracted from the homework data.

III.CONCLUSION

In this paper, we presented different case studies from in educational related data mining from prediction and classification techniques. Many researches show the potential of data mining in higher education. It was especially used to improve student's performance and detect early predictor of their Mark Based. Such Researches utilized the classification technique, decision tree in particular, to predict students' performance based on their grades on previous courses. For future work, we will generalize the study and add the elective to get more accurate process. We will extend the experiment using predictive data mining techniques.

REFERENCES

[1]. F. Castro, A.Vellido, A. Nebot, and F. Mugica, Applying data mining techniques to e-learning problems, In: L.C. Jain, R.A. Tedman, D.K. Tedman, (Eds.), Evolution of teaching and learning paradigms in intelligent environment. Studies in computational intelligence, Vol. 62, 2007, Springer, Berlin, Germany.

[2]. F. J. Mart´ınez, C. Herv´as, P. A. Guti´errez, A. C. Mart´ınez, and S. Ventura, Evolutionary Product-Unit Neural Networks for Classification, In: Conference on Intelligent Data Engineering and Automated Learning, 2006, pp 1320–1328.

[3]. Basma Boukenze1, Hajar Mousannif and Abdelkrim Haqiq, Predictive Analytics in Healthcare System Using Data Mining Techniques, DOI: 10.5121/csit.2016.60501

[4]. Parneet Kaura, Manpreet Singhb, Gurpreet Singh Josan, Classification and prediction based data mining algorithms to predict slow learners in education sector, 3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015)

[5]. Irina Ionita, Data Mining For Predicting the Military Career Choice, Revista Academiei Fortelor Terestre Nr. 3 (79)/2015.

[6]. B.Venkatalakshmi, M.V Shivsankar,Heart Disease Diagnosis Using Predictive Data mining, International Conference on Innovations in Engineering and Technology (ICIET'14)] , *Volume 3, Special Issue 3, March 2014* .

[7]. Soo-Yeon Ji, Sharad Sharma, Byunggu Yu, Dong Hyun Jeong, "Designing a Rule-Based Hourly Rainfall Prediction Model", IEEE IRI 2012, August – 2012.

[8]. M.Durairaj, K.Meena, ―A Hybrid Prediction System Using Rough Sets and Artificial Neural Networks‖ , International Journal Of Innovative Technology & Creative Engineering (ISSN: 2045-8711) VOL.1 NO.7 JULY 2011.

[9]. Hamidah Jantan, Abdul Razak Hamdan and Zulaiha Ali Othman, Human Talent Prediction in HRM using C4.5 Classification Algorithm, Hamidah Jantan et al. / (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2526-2534

[10]. R. Maciejewski et al., "Forecasting Hotspots - A Predictive Analytics Approach.," IEEE transactions on visualization and computer graphics, vol. 17, no. 4, pp. 440-453, May 2010.

[11]. Elia Georgiana Petre",A Decision Tree for Weather Prediction", Buletinul, Vol. LXI No. 1, 77-82, 2009.

[12]. Gerben W. Dekker, Mykola Pechenizkiy And Jan M. Vleeshouwers,

[13]. Z. Huang et al., "Managing Complex Network Operation with Predictive Analytics,"Proceedings of the AAAI Spring Symposium on Technosocial Predictive Analytics, *Science*, pp. 59-65, 2009.

[14]. R. M. Riensche et al., "Serious Gaming for Predictive Analytics," in AAAI Spring Symposium on Technosocial Predictive Analytics. Association for the Advancement of Artificial Intelligence (AAAI), San Jose, CA, no. Zyda, pp. 108-113, 2009.

[15]. Esther Ge, Richi Nayak, Yue Xu,Yuefeng Li Data Mining for Lifetime Prediction of Metallic Components, Copyright © 2006, Australian Computer Society, Inc, Vol. 61.

[16]. AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005) "Predicting survival time for kidney dialysis patients: a data mining approach", Elsevier Publication, Computers in Biology and Medicine ,Vol.35, pp 311–327

[17]. Behrouz Minaei-Bidgoli , Deborah A. Kashy , Gerd Kortemeyer , William F. Punch , "Predicting Student Performance: An Application Of Data Mining Methods With The Educational Web-Based System", LON-CAPA, 33rd ASEE/IEEE Frontiers in Education Conference, November 5-8, 2003, Boulder, CO.

